A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP6076.

Machine Learning-Based DNA Methylation Score for Fetal Exposure to Maternal Smoking: Development and Validation in Samples Collected from Adolescents and Adults

Sebastian Rauschert,[1] Phillip E. Melton,[2,3,4] Anni Heiskala,[5] Ville Karhunen,[6] Graham Burdge,[7] Jeffrey M. Craig,[8,9] Keith M. Godfrey,[10] Karen Lillycrop,[11] Trevor A. Mori,[12] Lawrence J. Beilin,[12] Wendy H. Oddy,[4] Craig Pennell,[13] Marjo-Riitta Järvelin,[5,6,14] Sylvain Sebert,[5,15] and Rae-Chi Huang[1]

[1]Telethon Kids Institute, University of Western Australia, Nedlands, Perth, Western Australia, Australia
[2]Centre for Genetic Origins of Health and Disease, University of Western Australia, Perth, Australia
[3]School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University, Perth, Australia
[4]Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania, Australia
[5]Center for Life Course Health Research, University of Oulu, Oulu, Finland
[6]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, Imperial College London, London, UK
[7]Institute of Developmental Sciences, University of Southampton, Faculty of Medicine, Southampton, UK
[8]Centre for Molecular and Medical Research, School of Medicine, Deakin University, Geelong, Victoria, Australia
[9]Molecular Epidemiology, Murdoch Children's Research Institute, Parkville, Australia
[10]MRC Lifecourse Epidemiology Unit and NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, UK
[11]Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, Hampshire, UK
[12]Medical School, Royal Perth Hospital Unit, University of Western Australia, Perth, Western Australia
[13]School of Medicine and Public Health, University of Newcastle, Newcastle, New South Wales, Australia
[14]Unit of Primary Care, Oulu University Hospital, Oulu, Finland
[15]Department of Metabolism, Digestion and Reproduction, Genomic Medicine, Imperial College London, London, UK

Background: Fetal exposure to maternal smoking during pregnancy is associated with the development of noncommunicable diseases in the offspring. Maternal smoking may induce such long-term effects through persistent changes in the DNA methylome, which therefore hold the potential to be used as a biomarker of this early life exposure. With declining costs for measuring DNA methylation, we aimed to develop a DNA methylation score that can be used on adolescent DNA methylation data and thereby generate a score for *in utero* cigarette smoke exposure.

Methods: We used machine learning methods to create a score reflecting exposure to maternal smoking during pregnancy. This score is based on peripheral blood measurements of DNA methylation (Illumina's Infinium HumanMethylation450K BeadChip). The score was developed and tested in the Raine Study with data from 995 white 17-y-old participants using 10-fold cross-validation. The score was further tested and validated in independent data from the Northern Finland Birth Cohort 1986 (NFBC1986) (16-y-olds) and 1966 (NFBC1966) (31-y-olds). Further, three previously proposed DNA methylation scores were applied for comparison. The final score was developed with 204 CpGs using elastic net regression.

Results: Sensitivity and specificity values for the best performing previously developed classifier ("Reese Score") were 88% and 72% for Raine, 87% and 61% for NFBC1986 and 72% and 70% for NFBC1966, respectively; corresponding figures using the elastic net regression approach were 91% and 76% (Raine), 87% and 75% (NFBC1986), and 72% and 78% for NFBC1966.

Conclusion: We have developed a DNA methylation score for exposure to maternal smoking during pregnancy, outperforming the three previously developed scores. One possible application of the current score could be for model adjustment purposes or to assess its association with distal health outcomes where part of the effect can be attributed to maternal smoking. Further, it may provide a biomarker for fetal exposure to maternal smoking. https://doi.org/10.1289/EHP6076

Introduction

Fetal exposure to maternal smoking during pregnancy increases the risk that the offspring will develop noncommunicable diseases (NCDs) (Agrawal et al. 2010; Bhattacharya et al. 2014; DiFranza et al. 2004; Hofhuis et al. 2003; Oken et al. 2005; Wakschlag et al. 2002; Wiklund et al. 2019). On average, 6% of the global female population are still smokers, although with a high degree of variability across countries (e.g., due to differences in the social and educational contexts, laws, and cultural factors), according to the WHO report on global tobacco epidemic 2017 (WHO 2017). A 2017 paper describing the smoking rates in Australia and Finland among other countries reported the smoking rates among young pregnant women are stagnant, despite the overall decrease in smoking rates (Reitan and Callinan 2017).

A meta-analysis by Oken et al. including 14 studies showed that offspring of mothers who smoked during gestation had a pooled adjusted odds ratio (OR) of 1.50 [95% confidence interval (CI): 1.36, 1.65] for the development of obesity (Oken et al. 2008). Timmermans et al. examined the association between maternal smoking during pregnancy and lower birth weight and the association between maternal smoking during pregnancy and higher weight gain and childhood overweight in the offspring (Timmermans et al. 2014). They showed that exposure to maternal smoking during pregnancy associated with an adjusted OR of 3.72 (95% CI: 1.33, 10.4) for the offspring being in the 85th BMI percentile (Timmermans et al. 2014).

The mechanisms through which maternal smoking may influence the health of the offspring have been suggested to involve the

Address correspondence to Sebastian Rauschert, Northern Entrance, Perth Children's Hospital, 15 Hospital Ave, Nedlands WA, Australia 6009, Telephone: +61863191589. Email:sebastian.rauschert@telethonkids.org.au

Supplemental Material is available online (https://doi.org/10.1289/EHP6076).

K.M.G. has received reimbursement for speaking at conferences sponsored by companies selling nutritional products and is part of an academic consortium that has received research funding from Abbott Nutrition, Nestec, Inc., and Danone S.A. The other authors declare they have no actual or potential competing financial interests.

Received 18 August 2019; Revised 20 August 2020; Accepted 28 August 2020; Published 15 September 2020.

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehponline@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

altered epigenetic regulation of genes. Epigenetics is the general term for changes to the DNA that are heritable through cell division and that relate to gene accessibility rather than DNA sequence changes (Goldberg et al. 2007). There are many different epigenetic mechanisms that can affect or alter gene accessibility, such as chromatin structural changes, histone modification or DNA methylation (Goldberg et al. 2007). Many studies have shown that maternal smoking during pregnancy is associated with highly reproducible and specific changes in differentially methylated cytosine-phosphate-guanine (CpG) base pairs in newborns (Joubert et al. 2016), children (Rzehak et al. 2016), young adults (Lee et al. 2015), and middle-age adults (Sun et al. 2013). In a meta-analysis with combined sample size of 6,685 newborns and 3,187 older children, 2,965 (FDR corrected, 568 after Bonferroni correction) differentially methylated CpGs in the offspring were associated with maternal smoking during pregnancy (Joubert et al. 2016). This included CpGs within *AHRR* (aryl-hydrocarbon receptor repressor), *MYO1G* (Myosin 1G), *CYP1A1* (Cytochrome P450 Family 1 Subfamily A Member 1), *GFI1* (Growth Factor Independent 1 Transcriptional Repressor) and *CNTNAP2* (Contactin-associated protein-like 2) (Rotroff et al. 2016; Rzehak et al. 2016; Tehranifar et al. 2018). These genes are associated with cancer development, detoxification of xenobiotics (*AHRR*) (Esser 2012), and adult body mass index (BMI) (*GFI1*) (Parmar et al. 2018) and suggest a possible epigenetic mechanism linking fetal exposure to maternal smoking during pregnancy with diseases in the offspring. Critically, Joubert et al. showed that the same CpGs were associated with fetal smoke exposure in cord blood, as well as in whole blood from 5-y-olds (Rzehak et al. 2016), and our own research suggests that fetal smoke exposure may induce persistent changes to the DNA methylome still detectable in middle age (Wiklund et al. 2019).

Machine learning is a subfield of artificial intelligence focusing on pattern detection. Machine learning methods can be divided into supervised and unsupervised methods. In supervised methods, labels are known, and the model tries to fit the data according to the label. In unsupervised methods, the algorithm tries to find clustering of similar data points. In both approaches, the aim is to create a model—with minimal assumptions on the data-generating process—that is generalizable to an external data set. Machine learning has proven to be useful in classification problems in medical research and diagnosis, especially in cancer and image classification (Capper et al. 2018; Díaz-Uriarte and Alvarez de Andrés 2006; Quraishi et al. 2015; Schmidhuber 2015; Yoo et al. 2014).

Successful examples of implementation of machine learning in epigenetics are Houseman's cell counts (Houseman et al. 2012) and Horvarth's epigenetic age acceleration (Horvath 2013). Both are widely adopted in the field and are based on the elastic net regression approach (Zou and Hastie 2005).

In light of that success, we applied machine learning methods to develop a DNA methylation score in adolescents and adults as a proxy for fetal exposure to maternal smoking. Similar to that described by Reese et al. (2017), we aimed to generate a score that could be applied to studies using HumanMethylation450K and EPIC BeadChip (Illumina) DNA methylation data. In comparison with the score by Reese et al. we have extended the DNA methylation score to older ages, including adolescence and adulthood. In data sets without information on maternal smoking during pregnancy, establishing and validating the score would enable its implementation in adjusting epigenome-wide DNA methylation association studies for this important early-life exposure. It may also serve as covariate to any model in more conventional epidemiological studies to adjust for possible confounding by maternal smoking in the absence of the measure. With reducing costs for DNA methylation arrays, the availability of such a DNA methylation score would be a valuable tool for epidemiological studies in disease pathways.

Methods

Studies

The Raine Study. The study design and initial characteristics of the Raine Study have been previously described (Newnham et al. 1993). From 1989 to 1991, a total of 2,900 pregnant women were enrolled. This included multiparous pregnancies. Recruitment took place at King Edward Memorial Hospital and surrounding private hospitals. The 2,868 live births have been followed up at 1, 2, 3, 5, 8, 10, 14 and 17 years during which anthropometric (e.g., height, weight, skinfolds), clinical, and biochemical data have been collected. Ethics approval for conducting the epigenetics analysis at the 17-y follow-up was given by the Human Ethics Committee of the University of Western Australia. Institutional ethics approval has been obtained through the University of Western Australia (approval numbers: RA/4/1/6613, 1214-EP, RA-4-1-2646). Informed and written consent was provided by the participants and their parents or legal guardians. The present analyses included 995 participants that were of white ethnicity.

The Northern Finland Birth Cohort 1986 (NFBC1986). The NFBC1986 consists of 99% of all children who were born in the provinces of Oulu and Lapland in northern Finland between 1 July 1985 and 30 June 1986 (Järvelin et al. 1993). There were 9,432 live-born individuals who entered the study. At the age of 16 y, those living still in Finland ($n = 9,215$) were invited to participate in a follow-up study, including a clinical examination. Overall, 7,344 participants attended the study in the year 2001/2002, of which 5,654 completed the postal questionnaire, completed the clinical examination, provided a blood sample, and gave written informed consent (parents and children). Approval for the studies was granted by the ethics committee of the Northern Ostrobothnia Hospital District in Oulu, Finland, in accordance with the Declaration of Helsinki.

The Northern Finland Birth Cohort (NFBC1966). The NFBC1966 is a prospective follow-up study of children from the two northernmost provinces of Finland (Rantakallio 1988). Of all women in this region with expected delivery dates in 1966, 96% were recruited through maternity health centers (12,058 live births). All individuals still living in northern Finland or the Helsinki area ($n = 8,463$) were contacted and invited for clinical examination when they turned 31 years of age (Järvelin et al. 2004). A total of 6,007 participants attended the clinical examination. DNA was extracted from blood samples given at the clinical examination (5,753 samples available). The samples were selected to resemble the original study cohort (Järvelin et al. 2004). An informed consent for the use of the data including DNA was obtained from all subjects and approval for the study was granted by the ethics committee of the Northern Ostrobothnia Hospital District in Oulu, Finland, in accordance with the Declaration of Helsinki.

DNA methylation profiling: the Raine Study. DNA methylation was measured in peripheral whole blood samples from participants at age 17 y using the Illumina Infinium HumanMethylation450K BeadChip. Venous blood samples were taken by phlebotomists after an overnight fast. Samples were stored at −80°C (176°F) until analysis. Processing of the Illumina Infinium HumanMethylation450K BeadChips was carried out by the Centre for Molecular Medicine and Therapeutics (CMMT) (http://www.cmmt.ubc.ca). We excluded three samples as outliers and one sample for biological sex inconsistency, because this might be indicative of a sample mix-up. Outliers were defined by the R packages *shinyMethyl* (version 1.22.0, Bioconductor) (Fortin and Hansen 2014) and *MethylAid* (version 1.22.0, Bioconductor) (Van Iterson et al. 2014) as samples that did not cluster together with the rest. Annotation of the CpG to the nearest gene was performed using

Illumina's genome coordinates (GRCh37/hg19). DNA methylation data of 996 white study participants on 475,429 probes were available for analysis.

DNA methylation profiling: NFBC1986. DNA was extracted from all 5,654 blood samples at the 16-y follow-up. DNA methylation was recorded on Illumina HumanMethlation450K array for 546 randomly selected subjects at the Department of Genomics Imperial College London (London, UK). Of those, 24 technical replicates were excluded. A total of 18 samples did not reach a call rate of >95% applying a detection *p*-value filter of 10×10^{-16}. We excluded seven samples with biological sex inconsistency, no sample was outlying from the overall data structure [first principle component (PC) score of the DNA methylation values outside mean ±4 standard deviations (SD)]. DNA methylation data of 517 samples from individuals of white ethnicity with 466,290 autosomal probes (call rate filter 95%) each were available for this analysis.

DNA methylation profiling: NFBC1966. DNA methylation at 31 years of age was measured for 807 randomly selected subjects of white ethnicity who attended the clinical examination and completed the questionnaire at both 31 and 46 years of age. For this, the Illumina HumanMethlation450K array was used at the Department of Genomics Imperial College London. For DNA methylation marker calling we used a detection *p*-value threshold of $<10^{-16}$. A call rate filter of 95% was applied to all autosomal Illumina probes, yielding 459,378 probes for association testing. Due to low marker call rate (<95%), 67 samples were excluded. Seven samples were excluded for biological sex inconsistency and one sample for globally outlying DNA methylation values (first PC score of the DNA methylation values outside mean ±4).

For all studies, we used the raw methylation betas without plate normalization because normalization methods might introduce bias into the model for the machine learning approach by changing the variance and residual structure of the DNA methylation data. This approach might also improve the application of the score to different data sets (Reese et al. 2017).

Smoking variables: the Raine Study. Mothers reported smoking behavior in questionnaires administered at the 18th and 34th week of gestation. Maternal smoking during pregnancy was coded as "yes" vs. "no" in regard to smoking during pregnancy, based on a combination of the categorical variables for the number of cigarettes smoked daily at 18 and 34 wk of gestation. In a previous epigenome-wide association study on the Raine Study data set, we did not find any differences between the CpGs associated with *in utero* smoke exposure at 18 wk and at 34 wk (Rauschert et al. 2019). To not sacrifice sample size, we decided to use the combined time points as any smoking, as described above.

We also present data on the number of adolescents that ever smoked. Smoking behavior of the adolescents at 16 years of age was self-reported in a confidential online questionnaire, and we recoded the variable asking for cigarette consumption over the lifetime to "(any) smoking" vs. "no smoking".

Figure S1 showing the distribution of exposure to maternal smoking across the plates for measurement of DNA methylation indicates no sign of a potential batch-induced bias.

Smoking variables: NFBC1986 and NFBC1966. Information on maternal smoking was self-reported in questionnaires by mothers during pregnancy. The questions asked and possible answers, respectively, were: Did you smoke before pregnancy? yes, no; Did you smoke when pregnancy was discovered? Yes, No; Number of cigarettes after the second gestational month: None, <10, 10 or more; Mother's smoking after the second gestational month: yes/no. This information was recoded to a binary variable indicating any smoking during pregnancy as opposed to no smoking during pregnancy to harmonize the data with the Raine Study variable.

The adolescents' own smoking in NFBC1986 and adults' smoking in NFBC1966 were also assessed using questionnaire data. We coded them into the category ever-smoker, which included occasional/former smoker and never smoker.

Analysis, Model Training, and Model Selection

Overview. A flow chart of the distinct modeling steps, including a brief description and which data was used is provided in Figure 1. The overall aim was to identify the best-performing algorithm to

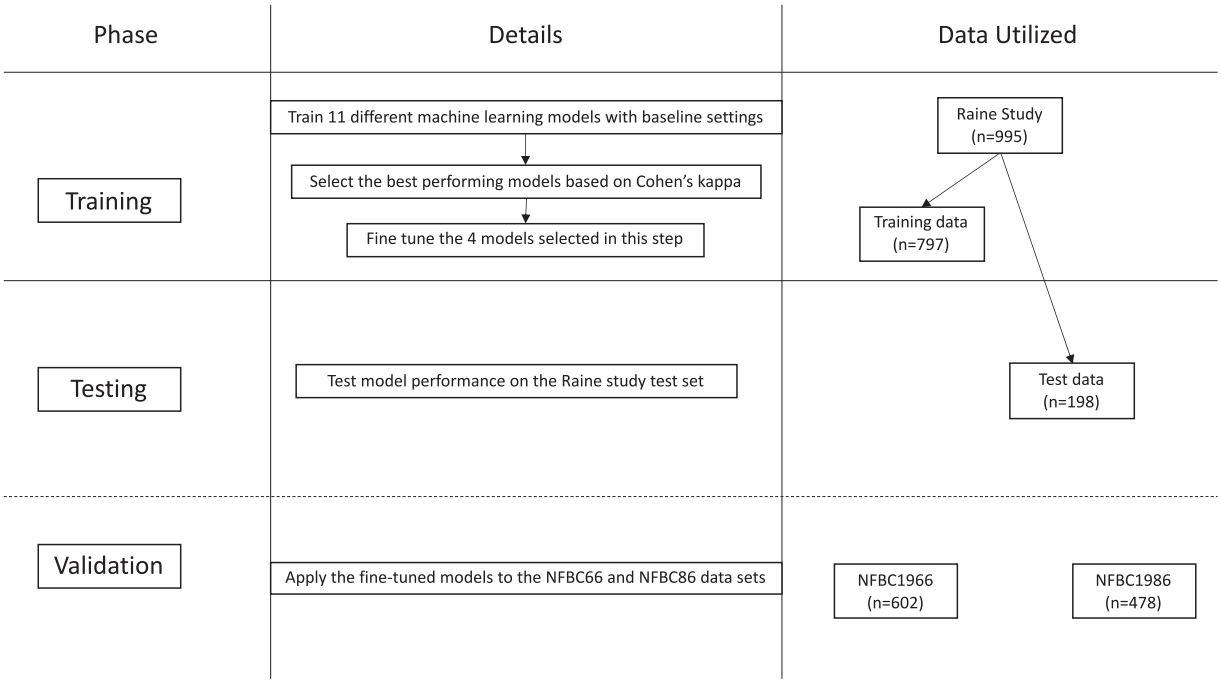

Figure 1. Flow chart for the modeling steps. This includes details of the steps undertaken in the training, testing and validation phase, as well as the data used per step.

identify study participants as being exposed to maternal smoking during pregnancy based on DNA methylation data. Performance in this context is defined as the model accuracy when compared with the known information of exposure to maternal smoking during pregnancy. To measure accuracy, we focused on Cohen's κ to identify the best model in this study. First, we split the Raine Study data into training and test set; then we applied 11 different machine learning algorithms to the training data to make a preselection of best-performing algorithms. The four best-performing models were taken forward for further refinement of the model parameters. Finally, the best-performing model was selected based on Cohen's κ.

Machine learning models. To check for the performance of different algorithms as defined in the previous paragraph, we derived scores for the exposure to maternal smoking during pregnancy from several different models after training with default settings for the model parameters in the statistical packages. The exact R code used for this can be found in Supplement S1. This was carried out to provide an overview of which methods to focus on for further modeling. There is no standard method of selecting machine learning algorithms. Some models are more suited for specific tasks than others, and a good way to start is to systematically test different algorithms on the data set and preselect those with the best initial performance for further training. It is advisable to select a variety of different algorithms, such as tree-, regression-, and clustering-based methods, because that allows for the testing of linear and nonlinear associations in the data.

All statistical and predictive modeling was conducted using R (version 3.5.1; R Development Core Team) and the *caret* package (Kuhn 2008). The primary models evaluated in this study were gradient boosting machine (Friedman 2001) using the *gbm* package (Ridgeway and Southworth 2013), elastic net regression (Zou and Hastie 2005) using the *glmnet* package (Hastie and Qian 2014), random forest (Breiman 2001) using the *randomForest* package (Liaw and Wiener 2002), and support vector machine (Cortes and Vapnik 1995) using *e1071* (Meyer and Wien 2015). In addition, we evaluated C5.0 (Pandya and Pandya 2015), Classification with Bagging (Breiman 1996), linear discriminant analysis (Duda et al. 2012), k-nearest neighbor (Altman 1992), naive Bayes classifier (Rish 2001), logistic regression, and classification and regression trees (Breiman et al. 1984), which were applied by setting the caret variable "*method*" to *C5.0*, *treebag*, *lda*, *knn*, *nb*, *glm*, and *rpart*, respectively. All the evaluated models other than k-nearest neighbor are supervised machine learning models.

Variable preselection. Overfitting in the variable selection when using the training data for that step was accounted for by selecting CpGs for the modeling process from the meta-analysis of Joubert et al. (2016). Joubert et al. used both FDR and the Bonferroni correction to define significant CpGs for their study. The table we selected the CpGs from is Supplement 4, Table S3 in Joubert et al., with the column titled "Meta-Analysis of sustained smoking and newborn methylation adjusted for cell type." We decided to consider an arbitrary $p < 0.00001$ for CpGs to be included in the modeling, which is in-between the FDR and Bonferroni threshold. We acknowledge that linear models can detect some relevant associations in the data but believe restricting the selection to only Bonferroni *p*-values limits the possibility to identify nonlinear associations using machine learning modeling. Including all $\sim$450,000 variables would technically be possible, but this would be computationally very expensive in terms of time and resources. Hence, our preselection process resulted in the inclusion of 1,511 CpGs.

As the aim of a predictive model is to be as parsimonious as possible, we excluded the highly correlated variables for the elastic net regression model, retaining only the CpGs that retain most of the information based on correlation structures of the data. This was done by examining the pairwise correlation structures of the CpGs in the Raine Study data before splitting it into training and test set. Given two CpGs were correlated with an $R^2 > 0.75$, we removed the CpG with the largest mean absolute correlation and thereby reduced multicollinearity issues. For this we used the R function *findCorrelation* (Kuhn 2008). In total, 267 CpGs were removed from the initial set of 1,511 CpGs because they had correlation coefficients >0.75 with at least one other CpG, leaving 1,244 CpGs for analysis (Table S1). The tree-based and support vector models are not as vulnerable to correlated data as the linear regression-based model; hence, all 1,511 CpGs were used for those.

Fitting method. We created the smoking score based on the study by Reese et al. using the exact coefficients and CpGs they identified with their LASSO approach (Reese et al. 2017). Briefly, to retrieve the score, one needs to multiply the CpG methylation values with the respective coefficient provided by Reese et al. in their supplement, Tables S1, and then add up the results from all 28 CpGs. The R code used to calculate the Reese score exemplified in the Raine Study, is provided in Supplement 1.

Richmond et al. describe two different scores for exposure to maternal smoking during pregnancy (Richmond et al. 2018). One score was created based on 568 CpGs from cord blood methylation data, and a second score used 19 CpGs from adult methylation data from the Avon Longitudinal Study of Parents and Children (ALSPAC) (Joubert et al. 2016; Richmond et al. 2018). Richmond et al. describe the calculation of the score as multiplying the model coefficients from the Joubert et al. meta-analysis of an EWAS (Epigenome Wide Association Study) for maternal smoking during pregnancy with the CpG methylation betas (Joubert et al. 2016). For that, the Supplement 4, Table S3, of the Richmond et al. publication is required. Hence, we identified the CpGs required for the score creation per Richmond et al. in the Raine Study, NFBC1966, and NFBC1986 and multiplied the individual participants DNA methylation values with the respective coefficient from the Joubert et al. study. For the 568 CpG score, the column titled "Meta-Analysis of sustained smoking and newborn methylation adjusted for cell type" is used, and for the 19 CpG score, the column "Meta-Analysis of sustained smoking and methylation in older children".

The random forest algorithm proposed by Breiman (2001) is a decision-tree–based algorithm. Rather than a single decision tree, this algorithm uses an ensemble approach. Every tree is created by only using a bootstrapped sample of the entire data. A second step of randomness is added for each split by selecting only a random subset of all predictive variables. This means, random forest implements both bagging (a method to combine multiple unstable learners, like decision trees, to gain more stable predictions) and random variable selection to build the trees, which leads to low correlation between the trees in the forest. For the purpose of this study, we set the *tuneLength* variable in the *caret* model to 20, which means a maximum of 20 different settings for the random forest parameters are evaluated. The following settings for the parameter *mtry* (number of input variables at each split) were tested: 2, 3, 5, 7, 10, 14, 19, 27, 38, 53, 74, 102, 142, 198, 275, 382, 530, 736, 1,023. For random forest, the *caret* package defaults the number of trees to 500 because the algorithm has been shown to plateau in its performance around this value.

The gradient boosting machine algorithm is also a tree-based method (Friedman 2002). Gradient boosting machines grow trees sequentially and try to improve on those trees that show weak predictions, making the method useful in cases of imbalanced data, as in this study. The parameters that can be tuned in a gradient boosting machine are minimum observations per node, number of trees and interaction depth [the number of splits to be

performed on a tree (starting from a single node)]. The following values and all their combinations were tested: All values from 1 to 20 for interaction depth, the minimum number of observations per node were kept at the *caret* default of 10, and the number of trees was tested in steps of 50 from 50 to 1,000.

The support vector machine (SVM) algorithm uses a subset of data points as so-called support vectors (Cortes and Vapnik 1995). In a two-dimensional case, support vectors are those data points that are closest to the line indicating the greatest separation between two classes. For the linear version of this model, the parameter *C* can be tuned. Also known as *Cost*, this parameter determines the possible misclassifications that are allowed. Simply speaking, it imposes an error penalty to the model. That means, the higher the value of *C*, in theory, the less likely it should be that the SVM algorithm will misclassify the data. In our tuning step, with *tuneLength* set to 20, the following values of *C* were tested: 0.00, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, and 5.00.

The code and the final models to use the three models above to create the score are available in Supplement 1 and require the *caret* package function *predict*.

The fourth model tested was elastic net regression (Zou and Hastie 2005). This model is a logistic regression-based model, that allows to specify two parameters: lambda and alpha. The model does not only fit the data, as in a logistic regression model, but it also performs variable selection. For this, the penalty parameter lambda can be tuned, which, based on its size, will penalize uninformative variables more. The alpha parameter can be set to 0, 1 or any integer in between, where 1 means LASSO regression, as in the Reese et al. study. There, the model strictly drops uninformative or correlated variables. The setting 0 for alpha does not perform variable selection but rather calculates weights for all variables, based on their importance for the classification. Elastic net regression keeps the alpha value between 0 and 1, which is a mix of both options described above; it will perform feature selection, but in the case of correlated variables that are both potentially meaningful for the classification, it will not randomly select one of the two, like LASSO.

With *tuneLength* of 20, the following parameter settings and all their combinations were tested: alpha of 0.10, 0.1473684, 0.1947368, 0.2421053, 0.2894737, 0.3368421, 0.3842105, 0.4315789, 0.4789474, 0.5263158, 0.5736842, 0.6210526, 0.6684211, 0.7157895, 0.7631579, 0.8105263, 0.8578947, 0.9052632, 0.9526316, and 1.0; lambda of 0.003785885, 0.004714173, 0.005870074, 0.007309400, 0.009101643, 0.011333339, 0.014112241, 0.017572522, 0.021881253, 0.027246473, 0.033927228, 0.042246086, 0.052604703, 0.065503224, 0.081564423, 0.101563783, 0.126466926, 0.157476249, 0.196088967, and 0.244169411.

For the final, best performing model using elastic net regression, the coefficients for the scoring are provided in Table S2, with instructions as to how to apply the score shown in Supplement 1. Briefly, the model can both produce a probability score (a value between 0 and 1, with 0 meaning not exposed and 1 meaning exposed) and a binary class (with a cutoff of 0.5; values above that fall into the "exposed" class, whereas values below fall into the "not exposed class"). To generate the DNA-methylation risk score (ranging from 0 to 1), the steps are as follows: *a*) multiply the CpG beta values by their respective coefficients generated by elastic net regression; and *b*) sum these across the 204 CpGs with the provided coefficients (Table S2).

We include a guide for easy application of this score (Supplement 1). This is taken from our R package, which is developed on github, so anyone can apply the score via the R programming language (https://github.com/Hobbeist/DNAsmokeR).

Of note, elastic net regression is the only machine learning model used in this study that not only fits a predictive model but also performs variable selection. This is why the creation of the score requires only 204 CpGs for elastic net regression, whereas all input CpGs are required for score creation in the other methods.

All models were trained using an 80% training and model-fitting subset and a 20% test subsample of the Raine Study data (Figure 1). The 80% training and model-fitting sample was determined using stratified randomized selection to retain the ratio between smoke exposure groups. For the training step in the 80% subset, we applied 10-fold cross-validation with 5 repeats. Therefore, the Raine model-fitting data set was randomly sampled 5 times into 10 groups, and for each sampling, 10 models were fitted (each time the model was fitted excluding one group, and predicted values were estimated for the remaining group). The average Cohen's κ results of those 50 modeling steps were used for comparing and selecting the best model. All CpG values were centered and scaled for the modeling. The R code for model training and testing of the above four machine learning models can be found in Supplement 1.

Imbalanced data problem. Classification algorithms try to reduce the overall error rate in classification; highly imbalanced data sets, where the minority class is very small in comparison with the majority class, tend to show good prediction accuracy but an overrepresentation of classification into the majority class. This also means, that prediction accuracy is not a feasible measure for overall classification quality. We used the following three approaches to address this: The data set was split into training and test data, stratified by exposure to smoking, meaning the ratio of smoke exposed to not exposed was the same in the training and test sets. To overcome the imbalance problem further, we applied a synthetic minority oversampling technique (Chawla et al. 2002), which outperforms oversampling the minority class (smoke exposed in our example) or undersampling the majority class (not exposed). In this approach, new, synthetic minority instances are created between existing data points, based on k-nearest neighbors.

We trained all our models on the kappa (κ) metric by Cohen (Cohen 1960; Viera and Garrett 2005). This metric compares the observed prediction accuracy with the expected prediction accuracy under random-guess circumstances:

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

where p_o is the observed prediction accuracy, and p_e is the expected prediction accuracy. For values between 0 and 1, 1 indicates a perfect prediction. We also report sensitivity, specificity, and area under the curve (AUC) because AUC has been established as a model comparison measure in machine learning (Held et al. 2016; Jin and Ling 2005).

Further, we report the Brier score as another quality measure, which is calculated by:

$$\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2,$$

where N is the number of forecasts, f_i is the score for participant i and o_i is the observed class: 0 for not exposed and 1 for exposed. The Brier score is a good quality measure for a probability score (Rufibach 2010). In terms of interpretation, values close to 0 indicate very good predictive power, whereas values close to 1 indicate bad performance.

Criteria for model selection. For the selection of the best model, the quality measure of interest was Cohen's κ (prediction accuracy and Cohen's κ for all models are reported in Figure S2).

The top four algorithms were elastic net regression, gradient boosting machine, SVM, and random forest. Hence, we decided to train those four models further and compare them based on Cohen's κ, sensitivity, specificity, and AUC. The model quality measures sensitivity, specificity, accuracy, and AUC were derived from the receiver operating characteristic (ROC) curve, using the point that minimizes the distance from the ROC curve to the top left corner, using the R packages *pROC*, *caret* and *plotROC*. The *caret* package was used to calculate the kappa statistic.

Final model exploration. **Model quality measures and variable importance.** For each machine learning model, we determined the CpGs that contributed the most toward the classification based on the absolute value of their estimated coefficients and compared the top 20 CpGs from each model to identify CpGs that were common across multiple models. To facilitate comparisons across models, we derived a measure of relative importance for each CpG by scaling the absolute value of each coefficient to the CpG coefficient with the largest value for each model.

This was done because these variables might hold insights into potential biological associations between maternal smoking during pregnancy and DNA methylation in the offspring.

Significance test for ROC differences. The elastic net regression–based smoking score, the Reese score, and the Richmond score are all continuous values that we inspect via ROC curves. To identify whether the observed difference in the ROC-curves and area under the ROC curves is statistically significant, we used DeLongs test. This test is used to check whether ROC curves are uncorrelated and is implemented in the R function *roc.test* (package *pROC*) (DeLong et al. 1988).

EPIC array sensitivity analysis. With the availability of the newer BeadChip array "EPIC" from Illumina, we also analyzed the performance of the score when only using the subset of CpGs in the elastic net model that are in both the 450k and EPIC array. In total, there are 23 elastic net score CpGs missing compared with 450k, totalling to 181 CpGs. We performed this sensitivity analysis only for the final best performing model, elastic net.

Results

Participants' characteristics (Table 1) showed no significant differences in the comparison of age, sex, adolescent smoking, and exposure to maternal smoking between the training ($n = 797$) and test sets ($n = 198$) of the Raine Study. Maternal smoking rates were similar in the training [ratio of exposed to not exposed: 0.42 (237/560)] and test sets [0.42 (59/139)], as expected due to the stratified split of the Raine Study into training and test sets. In both NFBC1986 and 1966, approximately 20% of study participants were exposed to maternal smoking during pregnancy. The NFBC cohorts had smoking rates of 34.3% at the 16-y follow-up (NFBC1986) and 52% at the 31-y follow-up (NFBC1966). These proportions of smokers is higher than in the Raine Study, where the proportion of smokers at the 17-y follow-up was 21.2% in the testing and 24% in the training set.

Machine Learning Models: Quality Measures

Taking into account that some CpGs available in the Raine Study were not available in the NFBC studies because of post-processing and outlier exclusion and because of the exclusion of highly correlated CpGs to create a sparse model, as well as variable selection via elastic net regression, the final number of CpGs to create the DNA methylation score was 204 for the elastic net (Table S2).

Scores Based on Gradient Boosting Machine, Random Forest, and SVM

For the gradient boosting machine approach, the cross-validation step resulted in a final model with 1,000 trees, an interaction depth of 6, and a minimum number of observations randomly selected per tree of 10.

The final SVM algorithm is a model with a penalization parameter C of 0.75. And finally, for the random forest: default number of trees (500) and *mtry* of 198 variables.

Final DNA Methylation Score

In the Raine Study data set, gradient boosting machine outperformed elastic net, random forest, and SVM scores on every measure except sensitivity, which was the same for gradient boosting machine and elastic net regression (Table 2). However, in the NFBC data sets, elastic net outperformed gradient boosting machine, random forest, and SVM scores for all quality measures except sensitivity (higher for gradient boosting machine than elastic net in all data sets, and for support vector machine in NFBC 1966). Based on this assessment, we concluded that elastic net regression, with an alpha of 0.1 and a lambda of 0.1264669, had the best overall performance of the four machine learning methods evaluated.

Table 1. Characteristics for the Raine study training and test data subset and the Northern Finland birth cohort 1986 and 1966.

	Raine Study: testing	Raine Study: training	*p*	NFBC1986	NFBC1966
n	198	797		478	602
Age (y) [mean (SD[a])]	17.20 (0.49)	17.27 (0.61)	0.132[b]	16.06 (0.36)	31.01 (0.34)
Sex (%)			0.975[c]		
Male	99 (50.0)	402 (50.4)		221 (46.2)	261 (43.4)
Female	99 (50.0)	395 (49.6)		257 (53.8)	341 (56.6)
Adolescent smoking (%)[d]			0.401[e]		
Non-smoker	108 (54.5)	392 (49.2)		288 (60.3)	283 (47.0)
Ever-smoker	42 (21.2)	191 (24.0)		164 (34.3)	313 (52.0)
Missing	48 (24.2)	214 (26.9)		26 (5.4)	6 (1.0)
Maternal smoking during pregnancy (%)[f]			1[c]		
Exposed	59 (29.8)	237 (29.7)		95 (19.9)	130 (21.6)
Not exposed	139 (70.2)	560 (70.3)		383 (78.7)	472 (78.4)

Note: *p* is the *p* value for the *t*-test and chi-square test between the Raine Study training and test set.
[a]SD: Standard deviation.
[b]Wilcoxon-Mann-Whitney *U*-test.
[c]Chi-square test.
[d]Adolescent smoking status was defined as ever smoked during the lifetime vs. never smoked as based on questionnaires.
[e]Wilcoxon Mann-Whitney *U*-test.
[f]Maternal smoking was defined as any smoking during pregnancy.

Table 2. Model quality measures (sensitivity, specificity, Cohen's κ, accuracy, AUC curve and Brier score) for the elastic net machine learning model, Reese et al. cord blood, Richmond et al. 568 CpG, Richmond et al. 19 CpG score the gradient boosting machine, random forest and support vector machine models that were among the four best performing models in our analysis. Results provided in this table are based on the Raine Study test data ($n = 198$), NFBC1986 ($n = 478$), and NFBC1966 ($n = 602$).

	Sensitivity	Specificity	Cohen's κ	Accuracy	AUC	Brier score	# CpGs required
Raine Study test data set							
Elastic net score	0.91	0.76	0.68	0.83	0.87	0.13	204
Gradient boosting machine	0.91	0.82	0.72	0.88	0.88	0.1	1,511
Random forest	0.87	0.73	0.58	0.83	0.83	0.17	1,511
Support vector machine	0.87	0.73	0.6	0.83	0.85	0.13	1,511
Reese score	0.88	0.72	0.6	0.83	0.85	0.21	28
Richmond score 568 CpGs	0.7	0.68	0.34	0.69	0.72	0.22	568
Richmond score 19 CpGs	0.79	0.58	0.37	0.72	0.73	0.22	19
NFBC1986							
Elastic net score	0.87	0.75	0.56	0.84	0.85	0.13	204
Gradient boosting machine	0.95	0.29	0.19	0.54	0.74	0.39	1,511
Random forest	0.79	0.16	0.06	0.64	0.54	0.24	1,511
Support vector machine	0.87	0.44	0.33	0.77	0.79	0.16	1,511
Reese score	0.87	0.61	0.46	0.82	0.8	0.18	28
Richmond score 568 CpGs	0.65	0.76	0.34	0.74	0.71	0.22	568
Richmond score 19 CpGs	0.65	0.77	0.31	0.68	0.73	0.22	19
NFBC1966							
Elastic net score	0.72	0.78	0.39	0.73	0.8	0.19	204
Gradient boosting machine	0.88	0.26	0.1	0.45	0.68	0.48	1,511
Random forest	0.77	0.18	0.05	0.64	0.48	0.24	1,511
Support vector machine	0.88	0.45	0.33	0.76	0.75	0.2	1,511
Reese score	0.72	0.7	0.32	0.71	0.73	0.18	28
Richmond score 568 CpGs	0.66	0.63	0.22	0.69	0.72	0.22	568
Richmond score 19 CpGs	0.61	0.72	0.23	0.63	0.73	0.22	19

Note: AUC, area under the receiver operator curve.

Previous "Gold Standard" Scores

Reese et al. score. The model evaluation metrics for the Reese score can be found in Table 2. For all metrics, including our key metric Cohen's κ, the elastic net regression–based score outperformed the Reese score in the Raine Study. However, when comparing the ROC curves (Figure 2), the regression elastic net–based curve did not significantly differ from the Reese score curve for the Raine Study based on the DeLong test (Table 3). In NFBC1986, the sensitivity is the same between the regression elastic net and the Reese score; however, all other model measures, including Cohen's κ, are better in the regression elastic net score. The DeLong test, as indicated in Table 3, shows a significant difference between the regression elastic net and the Reese score, with the elastic net–based score outperforming the Reese score. And the same is true for NFBC1966. There is no difference in sensitivity, but all other measures show that the regression elastic net score outperforms the Reese score.

Richmond et al. scores. For the Raine Study, the elastic net score outperforms the Richmond score with 568 and 19 CpGs in all model metrics. Further, as can be seen in Table 2, the ROC curves are significantly different between the elastic net and the Richmond based scores. The same is true for both NFBC studies, although the specificity is slightly better for the Richmond score with 568 CpGs in NFBC1986. Further, the Reese score outperforms both Richmond scores in all studies, except for the specificity in NFBC1986. Between the two Richmond scores, the 568 CpG score performs better than the 19 CpG score in all studies.

Variable importance. Seven CpGs were included in the top 20 CpGs for each of the four machine learning models, including the top four CpGs from the elastic net model: cg14179389 (*GFI1*, with the highest coefficient for the elastic net model), cg25949550 (*CNTNAP2*, 94% importance relative to cg14179389), cg22132788 (*MYO1G*, 80% relative importance), and cg11207515 (also in *CNTNAP2*, 66% relative importance) (Table S2). The remaining CpGs included in the top 20 CpGs for all machine learning models were cg13570656 (*CYP1A1*), cg17924476 (*AHRR*), and cg08474748 (*ANKRD31*) (Tables S3–S6).

For the gradient boosting machine algorithm, the first four CpGs were also frequently among the top CpGs in epigenome-wide association studies: cg22132788 (*MYO1G*, 100% importance), cg14179389 (*GFI1*, 99.99% importance), cg25949550 (*CNTNAP2*, 80.9% importance), and cg17924476 (*AHRR*, 20.8% importance).

The SVM algorithm identified CpGs in associated with *MYO1G*, *CNTNAP2*, *GFI1*, *CYP1A1*, *AHRR* and *FTO* among the top 10 most important variables in developing the score. Last, the random forest algorithm chose CpGs as the top 10 most important variables that are all frequently associated with maternal smoking during pregnancy. Those CpGs were associated with the genes *MYO1G*, *CNTNAP2*, *GFI1*, *AHRR*, *CYP1A1* and *FTO*.

EPIC array sensitivity analysis. The results for applying the elastic net score to those CpGs available in EPIC data (with 23 score CpGs missing compared with 450k, totaling to 181 CpGs for this analysis) can be found in Supplement 1, Table S1. For all metrics, the elastic net score based on the 450K version of Illumina's BeadChip array outperforms the elastic net score using only the 181 CpGs also available on the EPIC array. The overall performance, however, is still good, with Cohen's κ values all exceeding 0.3, with the best value in the Raine study being 0.65.

Discussion

In this study, we have developed a DNA methylation score for exposure to maternal smoking during pregnancy that outperforms an existing composite score (Reese et al. 2017), using DNA methylation probes (CpGs) measured in peripheral blood at the 17-y follow up of the Raine Study. We believe that with declining costs for measuring DNA methylation, such a DNA-methylation score could be a valuable contribution to epidemiological studies and clinical diagnostics.

To identify the most promising machine learning algorithms for creating the score, a range of models were chosen that

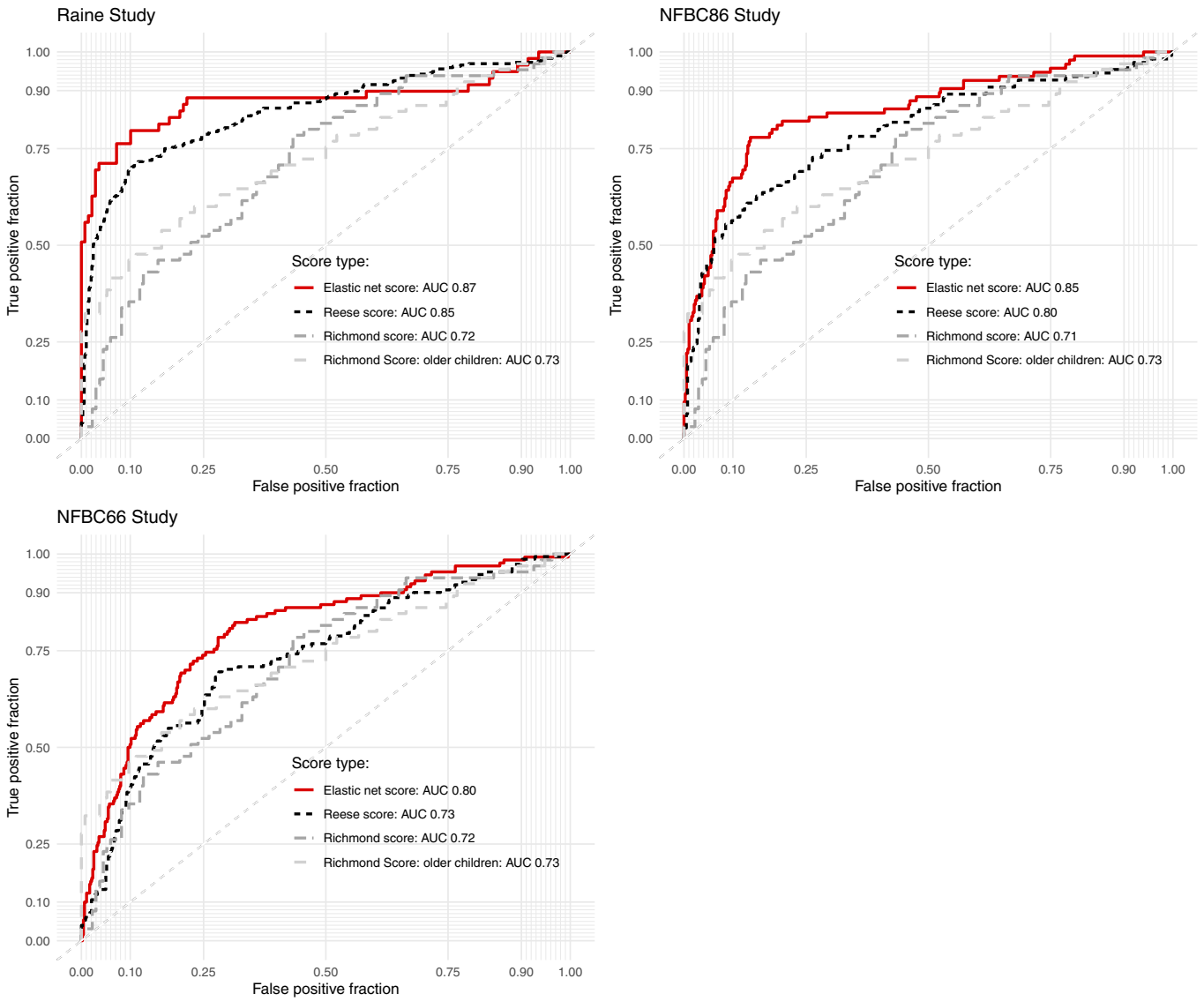

Figure 2. ROC for the four different model scores tested: elastic net regression, Reese et al. methylation score, Richmond et al. 568 CpG, and 19 CpG scores. AUC provided for every score, applied to the Raine Study test set, NFBC1986, and NFBC1966. Note: AUC, area under the ROC; ROC, receiver operator curve.

performed well in similar tasks and reflect a broad range of approaches from linear to nonlinear. We decided to further test the performance of elastic net regression, gradient boosting, SVM, and random forest algorithms, as they showed the best performance with respect to Cohen's κ in the algorithm selection step and are algorithms previously used in other epigenetic predictive modeling problems (Capper et al. 2018; Horvath 2013; Houseman et al. 2012).

Table 3. DeLong Test for significant difference between all ROC curves in Figure 2.

	DeLong Test *p*-values		
Test	Raine Study	NFBC1986	NFBC1966
Elastic net vs. Reese score	0.49	0.12	0.03
Elastic net vs. Richmond young	0.00058	0.008	0.006
Elastic net vs. Richmond old	0.01	0.04	0.004
Reese vs. Richmond young	1.57×10^{-6}	0.23	0.43
Reese vs. Richmond old	0.002	0.67	0.47
Richmond young vs. Richmond old	0.001	0.41	0.91

Our study aimed to establish a score using only variables created by the Illumina HumanMethylation450K BeadChip, because this makes the score independent of any other variables that studies otherwise would need to have collected. In comparison with more classical statistical approaches, the single aim of creating a score is that it is as accurate as possible in differentiating individuals exposed to *in utero* smoke from individuals not exposed to *in utero* smoke, with as little input as necessary.

Further, we purposely used raw DNA methylation betas, to avoid skewing the models based on normalization methods, which alter the residual and variance structure of the data. This approach also avoids the need for studies that aim to apply this score to normalize their data in a specific way, making it simple to apply. The main reason for doing so is that there is no gold standard of correcting the probes from the BeadChip, and the methods all perform slightly differently (Marabita et al. 2013). Ideally, the same normalization method would be applied to the training data as well as any data that applies the score. Using the raw values overcomes the issue initially and, as shown in this study, performs very well when applied to other studies.

In the Raine Study and NFBC1986, our score performs moderately better than the Reese et al. (2017) score created from cord blood. Reese et al. applied the LASSO penalized model rather than an elastic net approach to derive their score. Elastic nets have been mathematically shown to outperform LASSO regression when the number of variables is much larger than the number of cases (Zou and Hastie 2005). LASSO regression only selects at most as many variables as there are cases, which might not be feasible in the case of smaller sample sizes.

Nevertheless, the score by Reese et al. performs surprisingly well in the NFBC whole blood DNA samples collected at 16 and 31 years of age, with sensitivity consistently in excess of 70%, despite being derived based on cord blood DNA methylation. Because the cord blood measurement is closer to the exposure to maternal gestational smoking, the Reese et al. model might pick up stronger associations that potentially decrease over time, because DNA methylation has been shown to change with age (Horvath 2013).

The good predictive capability of our score applied to both the Raine test set and the validation study NFBC1986 (both 16 and 17 years of age) suggests that methylation could follow similar patterns across exposed and nonexposed individuals in the same age group and still holds some structural similarities when applied to a different age group, as in the NFBC1966.

Richmond et al. examined the relationship between maternal smoking and DNA methylation applying two different scores using 19 and 568 CpGs, and reported AUCs of 0.69 and 0.72, respectively, for their population of 656 women with measurements at two time points and in 230 men (Richmond et al. 2018). The two Richmond scores both underperformed the Reese et al. score and our own score when applied to the Raine data set and the two NFBC data sets. The CpGs were selected based on Bonferroni significance in association with maternal smoking, excluding the possibility that nonsignificant associations might still be contributing to the differences in association with maternal smoke exposure, by, for example, multivariate effects, that a linear regression model itself is not able to assess.

All four models trained in our study, independent of their performance, selected cg22132788 (*MYO1G*), cg25949550 (*CNTNAP2*), cg14179389 (*GFI1*), cg11207515 (*CNTNAP2*), cg13570656 (*CYP1A1*), cg17924476 (*AHRR*) and cg08474748 (*ANKRD31*) among the top 20 most influential variables for the DNA methylation score. These data concur with findings of several studies investigating DNA methylation in different age groups, showing that the same CpGs are differentially associated with [...] in utero exposure to maternal smoking (Joubert et al. 2014; Richmond et al. 2015; Rzehak et al. 2016).

The final best-performing overall score of this study, the elastic net, uses the *FTO* gene-associated cg00253658 for classification. The associated CpG is the 20th most important CpG based on the variable importance. This gene has previously been shown to associate with the development of obesity (Frayling et al. 2007). There is evidence, however, that single-nucleotide polymorphisms in the *FTO* gene might rather affect expression of the *IRX3* gene, which is related to obesity (Smemo et al. 2014). Further, methylation in the *AHRR* gene, with its associates CpG cg17924476 ranked sixth in the elastic net model, was associated with the development of eczema in boys and girls in a previous study (Mukherjee et al. 2016). Eczema in young children is a possible precursor of asthma and allergies, highlighting the potential association between exposure to maternal smoking during pregnancy and the development of asthma and allergies later in life (Almqvist et al. 2007).

This raises the possibility of using this score as a risk score for phenotypes associated with *in utero* smoke exposure such as obesity and allergic disease (Agrawal et al. 2010; DiFranza et al. 2004; Oken et al. 2005) in future studies that may give insights into pathways affected in fetal programming associated with maternal smoking.

Strengths and Limitations

DNA methylation is strongly associated with exposure to maternal smoking during pregnancy, as several studies have shown (Joubert et al. 2016; Rzehak et al. 2016). Hence, it is a good starting point to test which machine learning algorithms have the potential to be used as predictive models in the future.

With availability of DNA methylation data as measured by the Infinium HumanMethylation450K BeadChip and for all chosen models, the DNA methylation variables chosen for the classification, including the parameters for each model, are easily accessible and stated in this study. Because interpretability is important for reassurance when using in the context of clinical practice, we decided to approach the modeling with this in mind.

All maternal smoking variables were assessed via questionnaires rather than by the more objective measurement of cotinine, which is a limitation of this study.

Further, the score was developed using whole blood DNA methylation, which might not be the optimal sample type for specific DNA methylation and also might be affected by differences in white blood cell counts. We did not adjust for blood cell counts in our models, but our methylation score seemed to perform well despite this.

The performance of the score for correctly classifying exposure to maternal smoking during pregnancy might be influenced by current or recent smoking by adolescent or adult offspring, which might be more frequent in offspring exposed to maternal smoking during pregnancy. About half of the 31-y-old NFBC1966 study participants and over a third of the 16-y-old NFBC 1986 participants reported ever smoking. The proportion was lower in the Raine Study (21% in the training data set) but information on smoking was missing for almost one-quarter of the participants.

We used data from study participants of white ethnicity in training, testing, and validation data; hence, its performance needs to be confirmed in other ethnic groups. By using three different studies with two different age groups from culturally different countries (Australia and Finland), however, we were able to assess whether the models were overfit to the training study data or generalizable. The match between the Raine Study and NFBC1986 in terms of sex and follow-up is by chance, because both studies were independently developed, and data collection performed independently.

The model was created using data from the 450K version of the Illumina HumanMethylation series and not the newer EPIC version. The score still performed well when based only on the 181 CpGs available in the EPIC array using the coefficients in Table S2. Further, investigators whose data are missing other CpGs can also derive a score based on available CpGs, though the score's performance might differ, and the performance will be somewhat uncertain. We did not have pyro-sequenced data available to test how the smoking score would compare with the score derived from the 450K BeadChip array.

Further, although using raw DNA methylation data in our study showed very good results, future studies should systematically evaluate the effect of available normalization methods on the DNA-methylation score.

Conclusions

Our study shows that DNA methylation in late adolescence and early adulthood can be used to establish a score for the exposure

to maternal smoking during pregnancy. The score was validated externally in study populations from Finland and Australia.

We have evaluated the different machine learning approaches by using imbalanced data specific measures (Cohen's κ) as well as established comparative measures such as AUC, as in similar studies (Held et al. 2016; Jin and Ling 2005).

Our findings suggest that the score can be used by studies that have Illumina HumanMethylation450K or EPIC data available. As maternal smoking during pregnancy is one of the most well-established early life variables to be strongly associated with DNA methylation later in life, this score allows for studies that do not have information on maternal smoking behavior during pregnancy to account for its variance. For future studies, it might be interesting to test the interaction of this score with other risk factors related to maternal smoking during pregnancy.

The score combines information on DNA methylation and early-life exposure, and potentially a means to examine associations between this score and health outcomes such as cardiometabolic or respiratory diseases.

Acknowledgments

The authors thank the Raine Study participants and their families, the Raine Study Team for cohort coordination and data collection, The NHMRC for their long-term contribution to funding the study over the last 25 years, and The Telethon Kids Institute for long-term support of the Raine Study. We also acknowledge The University of Western Australia, Curtin University, Women and Infants Research Foundation, Edith Cowan University, Murdoch University, The University of Notre Dame Australia, and the Raine Medical Research Foundation for providing funding for core management of the Raine Study.

The authors further thank P. Rantakallio (launch of NFBC1966 and initial data collection). We gratefully acknowledge the contributions of the participants in the Northern Finland Birth Cohort 1966 study and the Northern Finland Birth Cohort 1986. The authors also thank all the field workers and laboratory personnel for their efforts.

This work was supported by resources provided by The Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

S.R. received funding for this work from the European Union's Horizon 2020 project LifeCycle, as an awardee of the LifeCycle fellowship 2018.

The DNA methylation work was supported by NHMRC grant 1059711. R.C.H. and T.A.M. are supported by NHMRC fellowships (grant number 1053384 and 1042255, respectively).

K.M.G. is supported by the UK Medical Research Council (MC_UU_12011/4), the National Institute for Health Research [as an NIHR Senior Investigator (NF-SI-0515-10042) and through the NIHR Southampton Biomedical Research Centre] and the European Union's Erasmus+ Capacity-Building ENeASEA Project and Seventh Framework Programme (FP7/2007-2013), projects EarlyNutrition and ODIN under grant agreement numbers 289346 and 613977.

NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital grant no. 2/97, 8/97; Ministry of Health and Social Affairs grant no. 23/251/97, 160/97, 190/97; National Institute for Health and Welfare Helsinki grant no. 54121; and Regional Institute of Occupational Health, Oulu, Finland grant no. 50621, 54231. NFBC1986 received financial support from EU QLG1-CT-2000-01643 (EUROBLCS) grant no. E51560; NorFA grant no. 731, 20056, 30167; USA/NIHH 2000 G DF682 grant no. 50945. S.S., A.H., V.K., and M.R.J. received support by H2020-633595 DynaHEALTH, H2020 733206 LifeCycle, the academy of Finland EGEA-project (285547), the Biocenter Oulu, H2020-733206 LifeCycle, H2020-824989 EUCANConnect, EU-H2020 (grant no. 82576), EU-H2020 EarlyCause (grant no. 848158), EU-H2020 LongITools (grant no. 873749), EU H2020-MSCA-ITN-2016 CAPICE Marie Sklodowska-Curie grant (grant no. 721567), and the Medical Research Council, UK [grant nos. MR/M013138/1, MRC/BBSRC MR/S03658X/1 (the EU JPI HDHL)]. V.K. is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (721567).

References

Agrawal A, Scherrer JF, Grant JD, Sartor CE, Pergadia ML, Duncan AE, et al. 2010. The effects of maternal smoking during pregnancy on offspring outcomes. Prev Med 50(1–2):13–18, PMID: 20026103, https://doi.org/10.1016/j.ypmed.2009.12.009.

Almqvist C, Li Q, Britton WJ, Kemp AS, Xuan W, Tovey ER, et al. 2007. Early predictors for developing allergic disease and asthma: examining separate steps in the 'allergic march.' Clin Exp Allergy 37(9):1296–1302, https://doi.org/10.1111/j.1365-2222.2007.02796.x.

Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. American Statistician 46(3):175–185, https://doi.org/10.2307/2685209.

Bhattacharya S, Beasley M, Pang D, Macfarlane GJ. 2014. Maternal and perinatal risk factors for childhood cancer: record linkage study. BMJ Open 4(1):e003656, PMID: 24394797, https://doi.org/10.1136/bmjopen-2013-003656.

Breiman L. 1996. Bagging predictors. Mach Learn 24(2):123–140, https://doi.org/10.1007/BF00058655.

Breiman L. 2001. Random forests. Machine Learning 45(1):5–32, https://doi.org/10.1023/A:1010933404324.

Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. *Classification and Regression Trees*. CRC Press/Taylor & Francis, Boca Raton, FL.

Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. 2018. DNA methylation-based classification of central nervous system tumours. Nature 555(7697):469–474, PMID: 29539639, https://doi.org/10.1038/nature26000.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. JAIR 16:321–357, https://doi.org/10.1613/jair.953.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46, https://doi.org/10.1177/001316446002000104.

Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn 20(3):273–297, https://doi.org/10.1007/BF00994018.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837–845, PMID: 3203132, https://doi.org/10.2307/2531595.

Díaz-Uriarte R, Alvarez de Andrés S. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:3, https://doi.org/10.1186/1471-2105-7-3.

DiFranza JR, Aligne CA, Weitzman M. 2004. Prenatal and postnatal environmental tobacco smoke exposure and children's health. Pediatrics 113:1007–1015, PMID: 15060193.

Duda RO, Hart PE, Stork DG. 2012. Pattern Classification. Hoboken, NJ: Wiley & Sons.

Esser C. 2012. Biology and function of the aryl hydrocarbon receptor: report of an international and interdisciplinary conference. Arch Toxicol 86(8):1323–1329, PMID: 22371237, https://doi.org/10.1007/s00204-012-0818-2.

Fortin JP, Fertig E, Hansen K. 2014. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Res 3:175, PMID: 25285208, https://doi.org/10.12688/f1000research.4680.2.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826):889–894, https://doi.org/10.1126/science.1141634.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Statist 29(5):1189–1232, https://doi.org/10.1214/aos/1013203451.

Friedman JH. 2002. Stochastic gradient boosting. Computat Stat Data Anal 38(4):367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. Cell 128(4):635–638, PMID: 17320500, https://doi.org/10.1016/j.cell.2007.02.006.

Hastie T, Qian J. 2014. Glmnet vignette. http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf [accessed 20 September 2016].

Held E, Cape J, Tintle N. 2016. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. BMC Proc 10(suppl 7):141–145, https://doi.org/10.1186/s12919-016-0020-2.

Hofhuis W, de Jongste JC, Merkus PJ. 2003. Adverse health effects of prenatal and postnatal tobacco smoke exposure on children. Arch Dis Child 88(12):1086–1090, https://doi.org/10.1136/adc.88.12.1086.

Horvath S. 2013. DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115, https://doi.org/10.1186/gb-2013-14-10-r115.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13:86, https://doi.org/10.1186/1471-2105-13-86.
Järvelin MR, Hartikainen-Sorri AL, Rantakallio P. 1993. Labour induction policy in hospitals of different levels of specialisation. Br J Obstet Gynaecol 100(4):310–315, PMID: 8494831, https://doi.org/10.1111/j.1471-0528.1993.tb12971.x.
Järvelin MR, Sovio U, King V, Lauren L, Xu B, McCarthy MI, et al. 2004. Early life factors and blood pressure at age 31 years in the 1966 Northern Finland Birth Cohort. Hypertension 44(6):838–846, https://doi.org/10.1161/01.HYP.0000148304.33869.ee.
Jin H, Ling CX. 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17:299–310, https://doi.org/10.1109/TKDE.2005.50.
Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. 2016. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. Am J Hum Genet 98(4):680–696, PMID: 27040690, https://doi.org/10.1016/j.ajhg.2016.02.019.
Joubert BR, Håberg SE, Bell DA, Nilsen RM, Vollset SE, Midttun O, et al. 2014. Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance? Cancer Epidemiol Biomarkers Prev 23(6):1007–1017, https://doi.org/10.1158/1055-9965.EPI-13-1256.
Kuhn M. 2008. Building predictive models in R using the caret package. J Stat Soft 28(5):1–26, https://doi.org/10.18637/jss.v028.i05.
Lee KWK, Richmond R, Hu P, French L, Shin J, Bourdon C, et al. 2015. Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environ Health Perspect 123(2):193–199, PMID: 25325234, https://doi.org/10.1289/ehp.1408614.
Liaw A, Wiener M. 2002. Classification and regression by randomForest. R News 2/3:18–22.
Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. 2013. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics 8(3):333–346, PMID: 23422812, https://doi.org/10.4161/epi.24008.
Meyer D, Wien FT. 2015. Support vector machines: the interface to libsvm in package e1071. https://cran.r-project.org/web/packages/e1071/e1071.pdf [accessed 14 February 2020].
Mukherjee N, Patil V, Chen S, Zhang H, Arshad SH, Holloway JW, et al. 2016. Interaction of *AHRR*-methylation and gestational smoking influences adolescent eczema, but not asthma. Eur Respiratory J 48(suppl 60):PA4594, https://doi.org/10.1183/13993003.congress-2016.PA459.
Newnham JP, Evans SF, Michael CA, Stanley FJ, Landau LI. 1993. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. The Lancet 342(8876):887–891, PMID: 8105165, https://doi.org/10.1016/0140-6736(93)91944-h.
Oken E, Huh SY, Taveras EM, Rich-Edwards JW, Gillman MW. 2005. Associations of maternal prenatal smoking with child adiposity and blood pressure. Obes Res 13(11):2021–2028, PMID: 16339135, https://doi.org/10.1038/oby.2005.248.
Oken E, Levitan EB, Gillman MW. 2008. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. Int J Obes (Lond) 32(2):201–210, https://doi.org/10.1038/sj.ijo.0803760.
Pandya R, Pandya J. 2015. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. Int J Comput Appl 117(16):18–21, https://doi.org/10.5120/20639-3318.
Parmar P, Lowry E, Cugliari G, Suderman M, Wilson R, Karhunen V, et al. 2018. Association of maternal prenatal smoking GFI1-locus and cardio-metabolic phenotypes in 18,212 adults. EBioMedicine 38:206–216, PMID: 30442561, https://doi.org/10.1016/j.ebiom.2018.10.066.
Quraishi BM, Zhang H, Everson TM, Ray M, Lockett GA, Holloway JW, et al. 2015. Identifying CpG sites associated with eczema via random forest screening of epigenome-scale DNA methylation. Clin Epigenetics 7:68, https://doi.org/10.1186/s13148-015-0108-y.
Rantakallio P. 1988. The longitudinal study of the northern Finland birth cohort of 1966. Paediatr Perinat Epidemiol 2(1):59–88, https://doi.org/10.1111/j.1365-3016.1988.tb00180.x.
Rauschert S, Melton PE, Burdge GC, Craig JM, Godfrey KM, Holbrook JD, et al. 2019. Maternal smoking during pregnancy induces persistent epigenetic changes into adolescence, independent of postnatal smoke exposure and is associated with cardiometabolic risk. Front Genet 10:770, https://doi.org/10.3389/fgene.2019.00770.
Reese SE, Zhao S, Wu MC, Joubert BR, Parr CL, Haberg SE, et al. 2017. DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. Environ Health Perspect 125(4):760–766, https://doi.org/10.1289/EHP333.
Reitan T, Callinan S. 2017. Changes in smoking rates among pregnant women and the general female population in Australia, Finland, Norway, and Sweden. Nicotine Tob Res 19:282–289, PMID: 27613884, https://doi.org/10.1093/ntr/ntw188.
Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. 2015. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet 24(8):2201–2217, PMID: 25552657, https://doi.org/10.1093/hmg/ddu739.
Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. 2018. DNA methylation as a marker for prenatal smoke exposure in adults. Int J Epidemiol 47(4):1120–1130, PMID: 29860346, https://doi.org/10.1093/ije/dyy091.
Ridgeway G, Southworth MH. 2013. Package 'gbm.' https://cran.r-project.org/web/packages/gbm/gbm.pdf [accessed 14 February 2020].
Rish I. 2001. An empirical study of the naive Bayes classifier. In: *Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3. New York, NY: IBM, 41–46.
Rotroff DM, Joubert BR, Marvel SW, Håberg SE, Wu MC, Nilsen RM, et al. 2016. Maternal smoking impacts key biological pathways in newborns through epigenetic modification in utero. BMC Genomics 17(1):976, https://doi.org/10.1186/s12864-016-3310-1.
Rufibach K. 2010. Use of Brier score to assess binary predictions. J Clin Epidemiol 63(8):938–939, PMID: 20189763, https://doi.org/10.1016/j.jclinepi.2009.11.009.
Rzehak P, Saffery R, Reischl E, Covic M, Wahl S, Grote V, et al. 2016. Maternal smoking during pregnancy and DNA-methylation in children at age 5.5 years: epigenome-wide-analysis in the European Childhood Obesity Project (CHOP)-study. PLoS One 11(5):e0155554, PMID: 27171005, https://doi.org/10.1371/journal.pone.0155554.
Schmidhuber J. 2015. Deep learning in neural networks: an overview. Neural Netw 61:85–117, PMID: 25462637, https://doi.org/10.1016/j.neunet.2014.09.003.
Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. 2014. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507(7492):371–375, PMID: 24646999, https://doi.org/10.1038/nature13138.
Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. 2013. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet 132(9):1027–1037, PMID: 23657504, https://doi.org/10.1007/s00439-013-1311-6.
Tehranifar P, Wu HC, McDonald JA, Jasmine F, Santella RM, Gurvich I, et al. 2018. Maternal cigarette smoking during pregnancy and offspring DNA methylation in midlife. Epigenetics 13(2):129–134, PMID: 28494218, https://doi.org/10.1080/15592294.2017.1325065.
Timmermans SH, Mommers M, Gubbels JS, Kremers SPJ, Stafleu A, Stehouwer CDA, et al. 2014. Maternal smoking during pregnancy and childhood overweight and fat distribution: the KOALA Birth Cohort Study. Pediatr Obes 9(1):e14–e25, PMID: 23362054, https://doi.org/10.1111/j.2047-6310.2012.00141.x.
Van Iterson M, Tobi EW, Slieker RC, Den Hollander W, Luijk R, Slagboom PE, Heijmans BT. 2014. MethylAid: visual and interactive quality control of large Illumina 450k datasets. Bioinformatics 23:3435–3437, PMID: 25147358, https://doi.org/10.1093/bioinformatics/btu566.
Viera AJ, Garrett JM. 2005. Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363, PMID: 15883903.
Wakschlag LS, Pickett KE, Edwin Cook J, Benowitz NL, Leventhal BL. 2002. Maternal smoking during pregnancy and severe antisocial behavior in offspring: a review. Am J Public Health 92(6):966–974, PMID: 12036791, https://doi.org/10.2105/ajph.92.6.966.
WHO (World Health Organization). 2017. WHO report on the global tobacco epidemic, 2017: Monitoring tobacco use and prevention policies. Geneva, Switzerland: World Health Organization.
Wiklund P, Karhunen V, Richmond RC, Parmar P, Rodriguez A, De Silva M, et al. 2019. DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. Clin Epigenet 11(1):97, PMID: 31262328, https://doi.org/10.1186/s13148-019-0683-4.
Yoo C, Ramirez L, Liuzzi J. 2014. Big data analysis using modern statistical and machine learning methods in medicine. Int Neurourol J 18(2):50–57, PMID: 24987556, https://doi.org/10.5213/inj.2014.18.2.50.
Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 67(2):301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.